Vo Declaration Exhibit I

EXHIBIT G

Data Review: libgen-fiction-books

Summary

- **Language:** The docs I looked at are written in coherent, well structured English covering a wide range of fictional settings.
- **Safety**: There was a surprising amount of NSFW material (mainly graphic sex and violence). I think a lot of it can probably be caught with some heuristic rules e.g. using a ban-list of NSFW words + thresholds.
- **Bad TOC**: Many of the documents have mangled or useless tables of contents (i.e. just chapter numbers, no titles); these could be caught with targeted regex rules.
- **Extraction errors**: Some documents had issues such as random, missing or misplaced whitespace, or interleaved page numbers or page titles that interfered with the coherence of the document. These seem like they'll be tricky to catch, but not sure how prevalent they are.
- **Copyright notices, legal disclaimers**: Many documents have copyright notices or legal disclaimers (e.g. "Any similarity ... is coincidental and not intended") near the beginning or end of the document. The notices are probably not ideal to include, and the disclaimers tend to use really similar boilerplate language which can be repetitive across many documents.
- **Promotional text**: Many include text toward the beginning or end promoting other titles by the same author, or the author's website, email, social media handles. The latter especially are probably not good to include.
- **About the author**: These include personal information about the author (e.g. familial details sometimes including names, their birth city, city of residence, etc). Obviously these are being volunteered by the author in a published work, but something to be aware of.

Raw notes

```
function read_doc() {
  doc_no=$1
  head -n$doc_no [redacted]libgen/fiction/fiction_en_deduped/fiction_e[redacted] | tail -n 1 | jq | sed 's/\\n/\n/g' | less -N
}
```

Doc 1:

- Bad TOC: `Chapter 1\nChapter 2 ...`
- Text bleed: "GET (5) FREE READS EVERY FRIDAY!"

Commented [1]: Thanks so much [redacted]@meta.com cc: [redacted]@meta.com David has some good observations on the fiction subset. Let's discuss what we can potentially do about these issues.

Commented [2]: [redacted]@meta.com, do we have the pipelines for unsafe content removal? Can we include LibGen datasets for the next runs? _Assigned to [redacted]@meta.com_

Commented [3]: We only explored toxicity so far; haven't updated the pipeline with new classifier, but we have a previous toxicity pipeline with [redacted]@meta.com ; this toxicity classifier could catch some of violence, but not sure about adult content

Commented [4]: I see, thx. I'm also not sure if we need to remove adult content at all

Commented [5]: Based on how they're trained the classifiers we've used (hatebert, toxdectroberta) are probably best at detecting abusive language as opposed to sexual or violent content.

The CRS says agents shouldn't generate sexually explicit content: https://docs.google.com/document/[redacted]

And we already do some NSFW filtering (e.g. the adult domain blocklist: https://github.com/fairinternal/[redacted]

However it's not clear to me either how far beyond that blocklist we should go during pre-training. I.e. is it better to filter entirely, or expose the model to it and fine-tune later...It's something worth testing in isolation if we do it

Commented [6]: yes, this we can clean out

Commented [7]: Currently we are removing all the copyright paragraphs from beginning and the end of the document. We can expand the list of disclaimers that should be removed.
1 total reaction

Commented [8]: [redacted]@meta.com, could we remove PII data from LibGen documents using your PII removal pipelines? _Assigned to [redacted]@meta.com_

Commented [9]: Ah the issue with the pipeline is that it detects any PII information within the whole document/book; Would we know if the authors typically put their info for example in the beginning/or end of the book? We could probably try to locate the section firs...

Commented [10]: I think we can use rule-based: first 10% of the book and last 10%, that should be enough.

- Safety: contains graphic scenes (sex, violence)

Doc 3:

- Odd whitespace…appears to have been an illustrated book
-
- Bad TOC: `1.\n2.\n ...`
- Copyright notices at the end

Doc 4:

- Extraction errors
  - Misplaced newlines
    - E.g. `"Now be nice to Willa Jean," said Mrs.\nQuimby, as…`
  - Page numbers left in
    - E.g. `"Ramona, 33\n\nwould you like …`
  - Missing whitespace
    - E.g. `she was a show-off and a nuisance.That hurt,`
- Copyright notices at the end

Doc 5

- Extraction errors
  - Repeated page titles left in
    - E.g. `So long, Nancy Drew, detective, she\n\nA Nancy Drew &> Hardy Boys SuperMystery\n\nthought, making her way out of her cabin and up to the Crown Deck. Nancy Drew, luxury-cruise passenger, is all ready for the sailing party to begin!`
    -

Doc 6

- Safety: graphic sexuality
- Bad TOC: `ONE\nTWO\nTHREE…`

Doc 7

- Safety: graphic sex and violence
- Legal disclaimer at the beginning: `The characters and events portrayed in this book are fictitious. Any similarity to real persons, living or dead, is coincidental and not intended by the author.`
- Copyright notices at the beginning

Doc 8

- Clean TOC and chapter headings, markdown formatted
- Contains references to notes section at the end

Doc 9

- No major issues
- Text bleed at the end: Visit thesameoldstreets.tumblr.com for free stories and extras!"

Doc 10

- Extraction errors
  - Every sentence split to a newline
  - Leading space with every full stop
- Promotional text at the end: `Facebook Fan Group: Micalea's Minions https://www.facebook.com/groups/ 370997336425131 /"`

Doc 11

- No major issues
- Extraction errors
  - Occasional extra whitespace: `C armen and Tommy walked down the street`

Doc 12

- No major issues
- Promotional text at the end (referencing other books)

Doc 13

- Bad TOC: `CHAPTER 1\n\nCHAPTER 2`
- Good chapter headings, markdown formatted
- Personal info about the author: `She now lives in Seattle, Washington, with her husband and two dogs`
- Promotional text + copyright notice at the end: `Visit Penguin.com for more about this author and a complete list of their books.`

Doc 14

- Extraction errors
  - Occasional extra whitespace: `P erhaps it was the wine talking,`

Doc 15

- Extraction errors
  - Extra whitespace (seems to impact all the capitalized section/chapter beginnings): `T HE POUNDING on the other side of the house had stopped`
- Promotional text + PII at the end: `Visit Rhys's blog at http://rhysford.wordpress.com/ or e-mail Rhys at` [HYPERLINK "mailto:rhys_ford@vitaenoir.com" \h] `."`

Doc 16

- Bad TOC: `Chapter One Chapter Two`
- Promotional text close to the end:
  - `Joel enjoys engaging with his readers. Drop by his website, www.joelgoldman.com , to find out more about him and his books. Read what he has to say about the writer's life on his blog, www.joelgoldman.com/blog and join him on Twitter at www.twitter.com/joelgoldman1 6300 and on Facebook at www.facebook.com/joelgoldmanauthor . Check out all his books at www.Amazon.com/author/joelgoldman . And watch videos about Joel and his books at www.youtube.com/user/joelgoldmanwriter .`
- No other major issues

Doc 17
- Safety: graphic sexual scenes
- Promotional text at the end

Doc 18
- No major issues

Doc 19
- Copyright notice at the beginning
- Safety: graphic sexual scenes
- Promotional text at the end: `For all titles by Diane Leyne, please visit\n\nSiren Publishing, Inc.\n\n`[HYPERLINK "http://www.sirenpublishing.com" \h]"

Doc 20
- Safety: graphic sexual scenes
- Potential PII: `TEXT LILLEY + YOUR EMAIL ADDRESS TO 16782493375 TO JOIN MY EMAIL NEWSLETTER.`
- Promotional text: `Visit my website for news and new releases here .`